ОСНОВЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ. РЕГРЕССИОННЫЙ И КОРРЕЛЯЦИОННЫЙ АНАЛИЗЫ

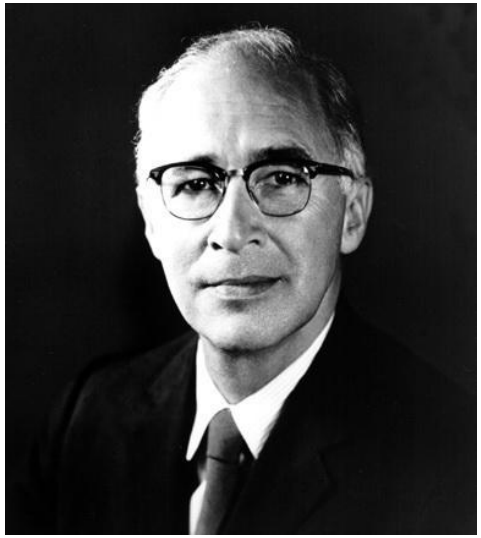
Лекция №3-4

к.т.н., доцент кафедры, Томин Н.В.

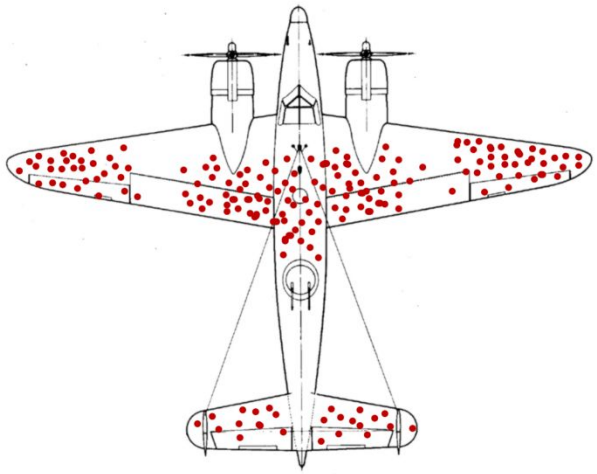
Основы математической статистики

Большой раздел современной математической статистики — статистический последовательный анализ, фундаментальный вклад в создание и развитие которого внес **А. Вальд** во время Второй мировой войны.

Систематическая ошибка выжившего – разновидность систематической ошибки отбора, когда по одной группе («выжившим») есть много данных, а по другой («погибшим») — практически нет. Так что исследователи пытаются искать общие черты среди «выживших» и упускают из вида, что не менее важная информация скрывается среди «погибших».



Абрахам Вальд,
венгерский математик и статистик.



Выборочный метод

Выборочный метод заключается в том, что из общей совокупности объектов, называемых **генеральной совокупностью**, извлекают некоторое число объектов, которое именуется выборкой. Эту выборку подвергают детальному исследованию, результаты которого можно применить ко всей генеральной совокупности. При выборочном методе исследуемый признак может быть распределён по «генеральной совокупности» неравномерно, поэтому выборка должна полностью отражать структуру генеральной совокупности.

Согласно теории вероятностей выборка будет правильно отражать свойства всей совокупности, если выбор производится случайно, т. е. так, что любая из возможных выборок заданного объема *n* из совокупности объема A имеет одинаковую вероятность быть фактически выбранной.

Пример использования «выборочного метода» – расчёт потерь

Пример.

Как правило, после **расчёта потерь электроэнергии в сетях 0,4 кВ** решается задача определения суммарных потерь в целом для сетей 0,4 кВ энергопредприятия на основании непосредственного расчёта потерь только в части из них. При этом, общее число линий обычно называют *генеральной совокупностью*, а рассчитываемую часть *выборкой*. Относительные потери электроэнергии в выборке с заданной доверительной вероятностью принимаются одинаковыми для всех сетей (*генеральной совокупности*) предприятия трансформаторных подстанций, находящихся на балансе предприятия.

Относительные потери электроэнергии в процентах для всей совокупности сети 0,4 кВ определяют по значениям (*K*) выбранных линий 0,4 кВ. Следует учесть, что для достоверных расчётов потерь в сети 0,4 кВ следует профессионально произвести объём выборки рассчитываемых линий, объединив сети в характерные группы с необходимым количеством воздушных и кабельных линий, а также близких и удалённых потребителей

Доверительные интервалы

Доверительный интервал – термин, используемый в математической статистике при интервальной оценке статистических параметров, более предпочтительной при небольшом объёме выборки, чем точечная. Доверительным называют интервал, который покрывает неизвестный параметр с заданной надёжностью.

Ежи Нейман,
польский и американский математик и статистик

Метод доверительных интервалов разработал американский статистик **Ежи Нейман**, исходя из идей английского статистика Рональда Фишера.

Суть метода заключается в следующем. По сделанной выборке *x1*, *x2*,⋯,*xn*, находятся числа x_{min} и x_{max} такие, чтобы выполнялось условие

$$P(x_{min} < \bar{x} < x_{max}) \geq p$$

Доверительные интервалы

Толкование доверительного интервала, основанное на интуиции, будет следующим:

Если уровень доверия *p* велик (скажем, 0,95 или 0,99), то доверительный интервал почти наверняка содержит истинное значение $\overline{x}$

Еще одно истолкование понятия доверительного интервала:

Его можно рассматривать как интервал значений параметра $\overline{x}$, совместимых с опытными данными и не противоречащих им.

Границы доверительного интервала определяются из следующих соотношений:

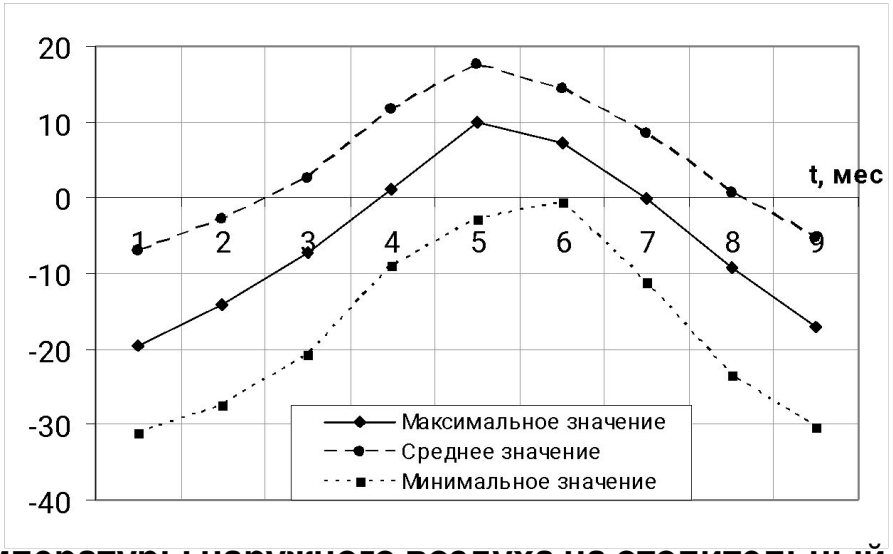
$$x_{min} = \bar{x} - \frac{\bar{\sigma}}{\sqrt{n}} t_\gamma; \qquad x_{min} = \bar{x} + \frac{\bar{\sigma}}{\sqrt{n}} t_\gamma,$$

где t_γ - квантиль нормального распределения.

Пример использования – планирование расхода тепловой энергии на основе прогноза температуры воздуха

Расход тепловой энергии для целей отопления в значительной степени предопределяется температурой наружного воздуха.

$$Q_{от} = q_{от} \cdot K \cdot T \qquad K = \frac{t_{вн} - t^{ф}_{н.в}}{t_{вн} - t^{р}_{н.в}}$$



Прогноз температуры наружного воздуха на отопительный период 2006 г. по Правобережному округу, г. Братска

Проверка статистических гипотез

Проверка статистических гипотез является содержанием одного из обширных классов задач математической статистики.

Статистическая гипотеза — предположение о виде распределения и свойствах случайной величины, которое можно подтвердить или опровергнуть применением статистических методов к данным выборки.

В ходе проверки статистических гипотез исследователь может столкнуться с возможностью допустить два вида ошибок:

1) отвергнуть правильную гипотезу – это **ошибка первого рода**;

2) принять неверную гипотезу – это **ошибка второго рода**.

Ошибку первого рода часто называют ***ложной тревогой***, ***ложным срабатыванием*** — например, анализ крови показал наличие заболевания, хотя на самом деле человек здоров, или металлодетектор выдал сигнал тревоги, сработав на металлическую пряжку ремня.

Ошибку второго рода иногда называют ***пропуском события*** или ***ложноотрицательным срабатыванием*** — человек болен, но анализ крови этого не показал, или у пассажира имеется холодное оружие,

Проверка статистических гипотез

Метод проверки статистической гипотезы состоит в следующем.

Производится выборка, на основе которой вычисляется значение ***t*** контрольной величины. Для проверки гипотез необходимо знать контрольную величину функции ***T*** от рассматриваемой выборки, меньше значения которой гипотеза будет считаться неверной. Если вероятность события ***t<T*** меньше уровня максимально допустимой вероятности ошибки первого рода (отвергнуть правильную гипотезу), то гипотеза принимается.

Критерий принятия условием $P(t < T) < \alpha$

В соответствии со стандартными подходами, принятыми в теории стат. гипотез, мерой надежности является уровень значимости принятия гипотезы: величина ***α***. Чем меньше а, тем «осторожнее» гипотеза. Надежность принятия гипотезы (1−***α***.) для

Проверка статистических гипотез

1. **Критерии значимости** – *проверка гипотез о нормальности выборки*
   Критерий Шапира-Уилка
   Критерий хи-квадрат Пирсона и др.
2. **Критерии согласия** – *это критерии проверки гипотез о соответствии эмпирического распределения теоретическому распределению вероятностей.*
   Критерий Колмогорова-Смирнова
   Критерий согласия хи-квадрат Пирсона
   и др.
3. **Критерии однородности** – *это критерии проверки гипотез о том, что две (или более) выборки взяты из одного распределения вероятностей.*
   Критерий Стьюдента
   Критерий Фишера-Снедекора
   Критерий Кохрена и др.

Пример использования – Бомбардировка Лондона

Пример. Задача о бомбардировках Лондона. Задача возникла в связи с бомбардировками Лондона во время Второй мировой войны. Для улучшения организации оборонительных мероприятий, необходимо было понять цель противника. Для этого территорию города условно разделили сеткой из 24-х горизонтальных и 24-х вертикальных линий на 576 равных участков. В течении некоторого времени в центре организации обороны города собиралась информация о количестве попаданий снарядов в каждый из участков. В итоге были получены следующие данные:

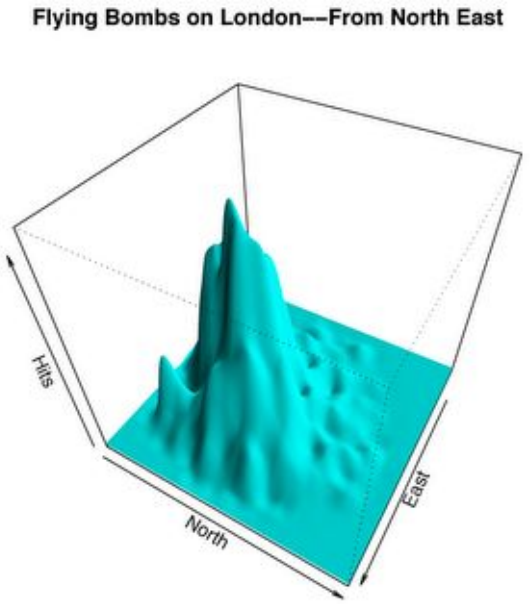
Число попаданий	0	1	2	3	4	5	6	7
Количество участков	229	211	93	35	7	0	0	1

Пример использования – Бомбардировка Лондона

Как видно из графика распределение сброшенных на Лондон бомб далеко от равномерного закона, но является ли это свидетельством точного нацеливания?

Всего 537 бомб упали на 576 квадрата, что около одной бомбы на квадрат в среднем. Исследователи подставили эти числа в формулу Пуассона, чтобы узнать сколько скоплений ожидается получить случайным образом

Количество бомб на квадрат	Предполагаемое кол-во квадратов по Пуассону	Действительное кол-во квадратов
0	226·74	229
1	211·39	211
2	98·54	93
3	30·62	35
4	7·14	7
5 и больше	1·57	1
	576·00	576



Визуализация количества бомб, сброшенных над различными частями города

Пример использования – Бомбардировка Лондона

Гипотеза H0: стрельба случайна (нет "целевых" участков).

$P\{S=j\}=\frac{\lambda^j}{j!}e^{-\lambda}$, где S - число попаданий, $\widehat{\lambda}=0.924$ (она)

$$\chi^2=\sum_{j=1}^{k}\frac{\left(n_j-E_j\right)^2}{E_j}=32.6\sim\chi^2_{8-1-1}$$

Тогда при уровне значимости 0.05 гипотеза H0 не выполняется

Объединим события (4,5,6,7) с малой частотой попаданий в одно (поправка Йетса), тогда имеем:

Число попаданий	0	1	2	3	4-7
Количество участков	229	211	93	35	8

$$\chi^2=1.05\sim\chi^2_{5-1-1}$$

тогда при 0.05 **гипотеза H0 всё-таки верна.**

Отсев грубых ошибок

Исходные данные, получаемые в результате экспериментов, в силу разных причин, могут содержать грубые ошибки или аномальные наблюдения, которые должны быть исключены из выборочной совокупности.

Наиболее простой метод отсева грубых ошибок при нормальном законе распределения – использование **правила трех сигм**, которое формулируется следующим образом: *разброс случайных величин от их среднего значения не должен превышать трех среднеквадратичных отклонений.*

Алгоритм отсева грубых ошибок состоит в следующем.

1. Рассчитываются выборочное среднее и среднеквадратичное отклонение.
2. Вычисляются значения *Xmax* и *Xmin* по формулам

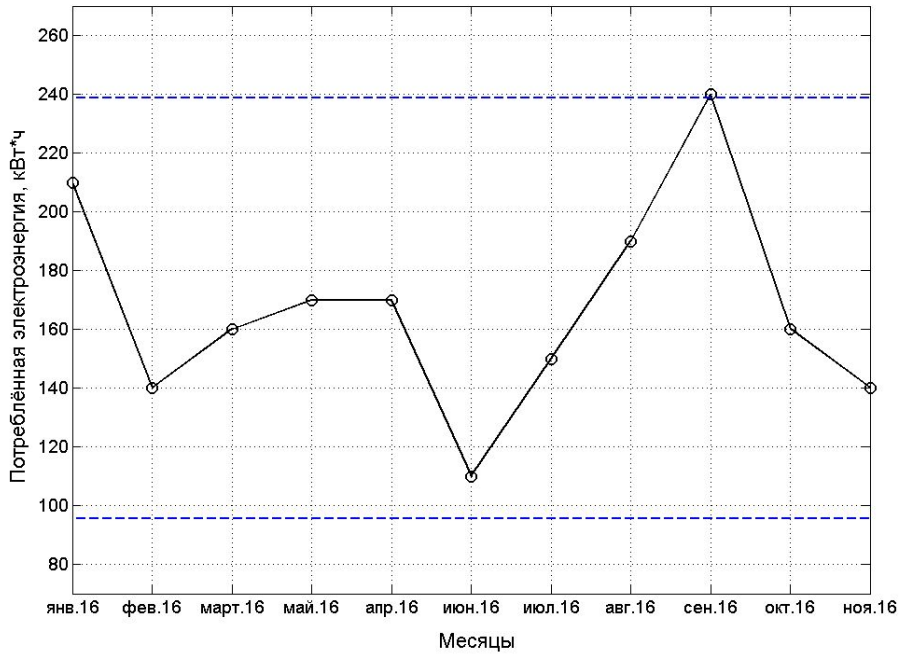
$$x_{min} = \bar{x} - 3\bar{\sigma}, \;\; x_{max} = \bar{x} + 3\bar{\sigma}$$

3. Величины, находящиеся за пределами интервала (+ ; +) исключаются из выборочной совокупности как недостоверные или аномальные и обработка результатов эксперимента по определению точечных оценок производится повторно.

Пример использования – Аномальные значения в электропотреблении

Пример. Имеются данные потребления электроэнергии в жилой квартире в период с января по ноябрь. Необходимо установить не содержат данные показания грубых ошибок.

Янв.	Февр.	Март	Апр.	Май	Июнь	Июль	Авг.	Сент.	Окт.	Нояб.
210	140	160	170	170	110	150	**190**	**240**	160	140

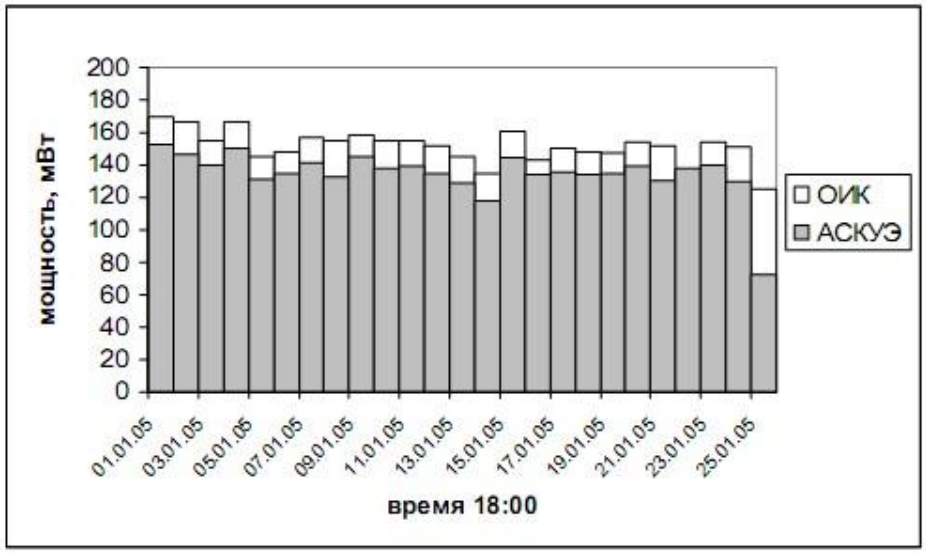


Пример использования – Достоверизация телеизмерений мощности

Имея выборку, состоящую из телеизмерений (ТИ) перетока мощности по данным ОИК и АСКУЭ за некоторый интервал времени, для решения задачи достоверизации ТИ перетока мощности можно использовать метод Стьюдента, которые считается связанным с нормальным распределением, что соответствует закону распределения ошибок в ТИ.

Учитывая это, процедуру отсева грубых погрешностей измерений можно представить в виде примера, где выборка представляет из себя ретроспективные данные о перетоке мощности по межсистемной ВЛ–220 кВ

$$\tau_{(p,n)} = \frac{t_{(p,n-2)}\sqrt{(n-1)}}{\sqrt{n-2+[t_{(p,n-2)}]^2}}.$$



Пример использования – Достоверизация телеизмерений мощности

Выходная форма контроля достоверности измерений перетоков мощности.

ОИК, МВт	АСКУЭ, МВт	Δ, МВт	τ	заключение
170,00	152,70	17,30	0,11	БЕЗ ОТСЕВА
---//---				
157,00	141,77	15,23	0,58	БЕЗ ОТСЕВА
155,00	132,85	22,16	1,74	ОТСЕВ по усмотрению
159,00	144,94	14,06	0,98	БЕЗ ОТСЕВА
---//---				
161,00	144,41	16,59	0,13	БЕЗ ОТСЕВА
143,00	133,95	9,05	2,66	ОТСЕВ по усмотрению
150,00	135,38	14,62	0,79	БЕЗ ОТСЕВА
---//---				
152,00	130,47	21,53	1,53	БЕЗ ОТСЕВА
136,00	137,44	1,44	5,21	БЕЗ ОТСЕВА
154,00	139,76	14,24	0,92	БЕЗ ОТСЕВА
151,00	129,94	21,06	1,37	БЕЗ ОТСЕВА
125,00	72,39	52,61	11,96	ОТСЕВ
Среднее		16,97		
Дисперсия		8,88		
СКО		2,98		

Подобный метод достове-ризации ТИ мощности поз-воляет контролировать тех-ническое состояние систем сбора данных, эффективно выявлять грубые ошибки ТИ

Регрессионный анализ

Величины, характеризующие различные свойства объектов, могут быть независимыми или взаимосвязанными. Различают два вида взаимосвязи: **функциональную** и **статистическую.**

В реальных ситуациях существует бесконечно функциональные связи являются математическими абстракциями. В реальности многие параметры следует считать случайными, что исключает проявление однозначного соответствия значений. Воздействие общих факторов, наличие объективных закономерностей в поведении объектов приводят лишь к проявлению статистической зависимости.

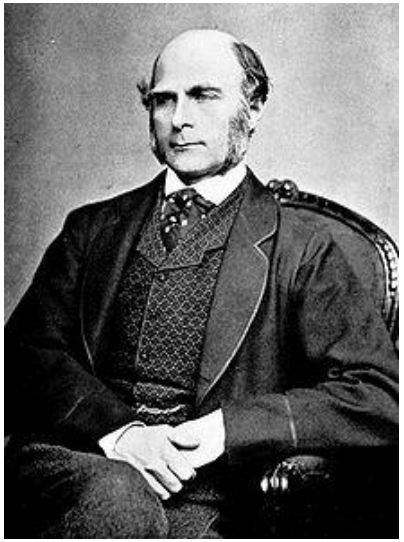
Статистической называют зависимость, при которой изменение одной из величин влечет изменение распределения других (другой), и эти другие величины принимают некоторые значения с определенными вероятностями.

Более важным частным случаем статистической зависимости является **корреляционная зависимость**, характеризующая взаимосвязь значений одних случайных величин со средним значением других, хотя в каждом отдельном случае любая взаимосвязанная величина может принимать различные значения. Если же у взаимосвязанных величин вариацию имеет только одна переменная, а другая является детерминированной, то такую связь называют не корреляционной, а **регрессионной.**

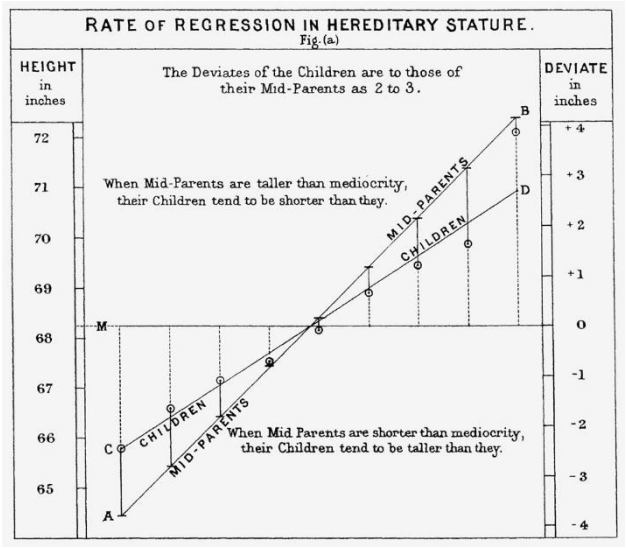
Регрессионный анализ

Термин **"регрессия"** был введён Фрэнсисом Гальтоном в конце 19-го века. Гальтон обнаружил, что дети родителей с высоким или низким ростом обычно не наследуют выдающийся рост и назвал этот феномен "регрессия к посредственности« (или регрессия к среднему). Он наглядно объяснил, что рост детей усредняется относительно роста родителей.

К примеру, независящие от человека ситуации также подвергаются регрессии. Механизм заключается в следующем: все пиковые ситуации, достигнув максимальной отметки, начинают откатываться к среднему состоянию.



Сэр Фрэнсис Гальтон, английский исследователь, географ, антрополог и психолог.



Регрессионный анализ

Регрессия — зависимость математического ожидания (например, среднего значения) случайной величины от одной или нескольких других случайных величин (свободных переменных), т.е.

Регрессионным анализом называется поиск такой функции ***f,*** которая описывает эту зависимость. Регрессия может быть представлена в виде суммы неслучайной $M(y|x) = f(x)$ случайной составляющих.

В общем случае для стандартизованных данных функциональную зависимость показателя от параметров можно представить в виде

$$y = f(x_1, x_2, \cdots, x_m) + e$$

где *f* – заранее не известная функция, подлежащая определению;

e – ошибка аппроксимации данных.

Регрессионный анализ

В целях выбора функциональной связи заранее выдвигают гипотезу о том, к какому классу может принадлежать функция ***f***, а затем подбирают "лучшую" функцию в этом классе. Выбранный класс функций должен обладать некоторой "гладкостью", т.е. "небольшие" изменения значений аргументов должны вызывать "небольшие" изменения значений функции (*обычно данные содержат некоторые ошибки измерений, а само поведение объекта (к примеру, микропроцессорных реле) подвержено влиянию помех, маскирующих истинную связь между параметрами и показателем*).

Простым, удобным для практического применения и отвечающим указанному условию является класс полиномиальных функций

$$y = a_o + \sum_{j=2}^{m} a_j x_j + \sum_{j=2}^{m-1} \sum_{k=j+1}^{m} a_{jk} x_j x_k + \sum_{j=2}^{m} a_{jj} x_j^2 + \cdots + e$$

Частным случаем, широко применяемым на практике, является полином первой степени или уравнение линейной регрессии

$$y = a_o + \sum_{j=2}^{m} a_j x_j + \mathrm{e}$$

Регрессионный анализ

Линейная регрессия

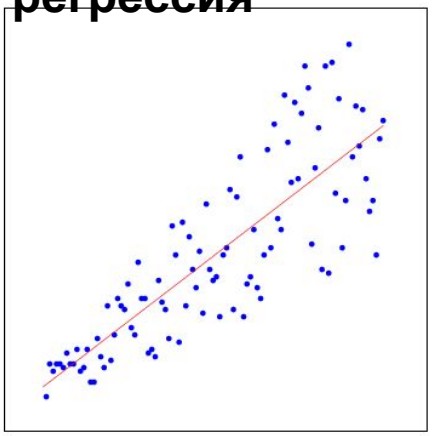


Рис. 1: линейная регрессия: аналитический метод

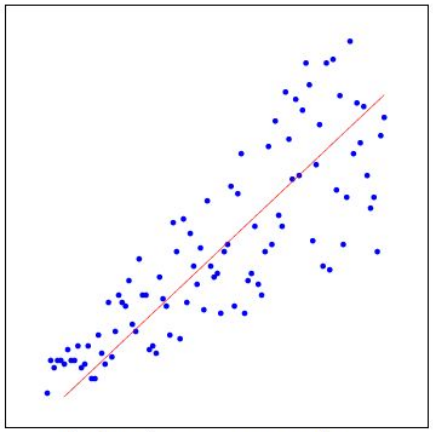


Рис. 2: линейная регрессия: градиентный спуск

Нелинейная регрессия

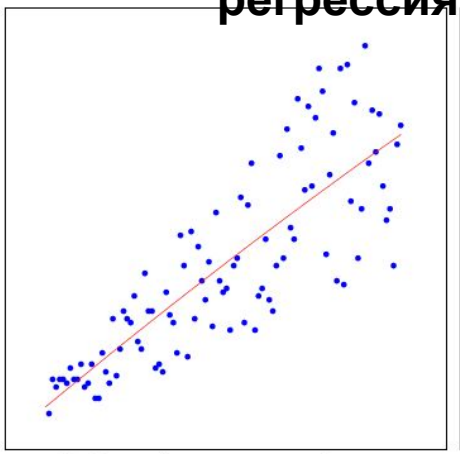


Рис. 3: нелинейная регрессия: полином 2 степени

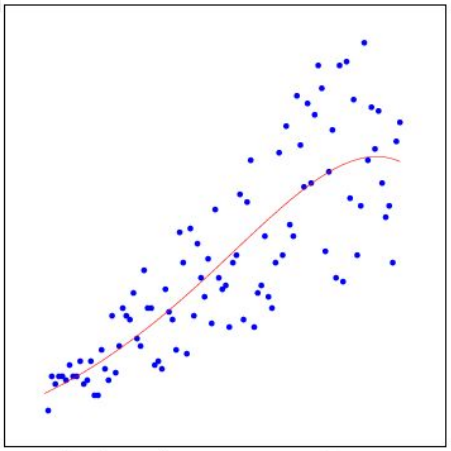


Рис. 4: нелинейная регрессия: полином 5 степени

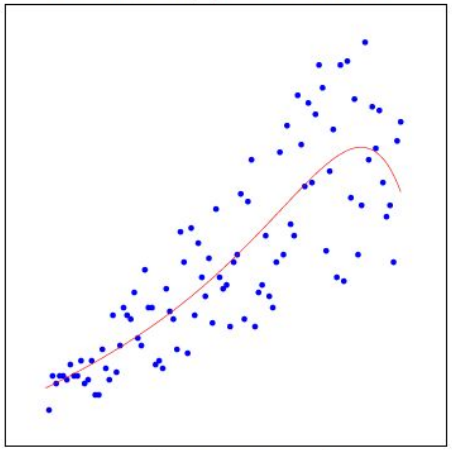


Рис. 5: нелинейная регрессия: полином 12 степени

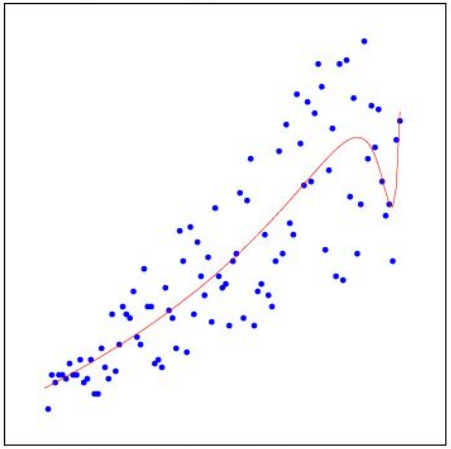


Рис. 6: нелинейная регрессия: полином 72 степени

Регрессионный анализ

Решение задачи регрессионного анализа целесообразно разбить на несколько этапов:

- предварительная обработка данных;
- выбор вида уравнений регрессии;
- вычисление коэффициентов уравнения регрессии;
- проверка адекватности построенной функции результатам наблюдений.

Регрессионный анализ проводится при следующих допущениях:

1. количество наблюдений достаточно для проявления статистических закономерностей относительно факторов и их взаимосвязей;
2. обрабатываемые данных содержат некоторые ошибки (помехи), обусловленные погрешностями измерений, воздействием неучтенных случайных факторов;
3. матрица результатов наблюдений является единственной информацией об изучаемом объекте, имеющейся в распоряжении перед началом исследования.

Регрессионный анализ

Уравнение регрессии в регрессионном анализе следует трактовать как векторное, ибо речь идет о матрице данных.

При этом, полученную систему уравнений на основе имеющихся данных однозначно решить невозможно, так как количество неизвестных всегда больше количества уравнений.

$$\begin{cases} a_0 + a_1 \cdot x_0 + a_2 \cdot x_0^2 + \cdots + a_n \cdot x_0^n + e_0 = y_0, \\ a_0 + a_1 \cdot x_1 + a_2 \cdot x_1^2 + \cdots + a_n \cdot x_1^n + e_1 = y_1, \\ \cdots\cdots\cdots\cdots \\ a_0 + a_1 \cdot x_n + a_2 \cdot x_n^2 + \cdots + a_n \cdot x_n^n + e_n = y_n. \end{cases}$$

Здравый смысл подсказывает: ***желательно выбрать коэффициенты полинома так, чтобы обеспечить минимум ошибки аппроксимации данных.*** Могут применяться различные меры для оценки ошибок аппроксимации. В качестве такой меры нашла широкое применение **среднеквадратическая ошибка.**

Регрессионный анализ – метод наименьших квадратов

На ее основе разработан специальный метод оценки коэффициентов уравнений регрессии – **метод наименьших квадратов (МНК),** когда минимизируется сумма квадратов отклонений реально наблюдаемых Y_k от их оценок $\hat{Y}_k$ (*имеются в виду оценки с помощью прямой линии, претендующей на то, чтобы представлять искомую регрессионную зависимость*):

$$\sum_{k=1}^{M}(Y_k - \hat{Y}_k)^2 \to min$$

Для решения задачи регрессионного анализа МНК вводится понятие **функции невязки:**

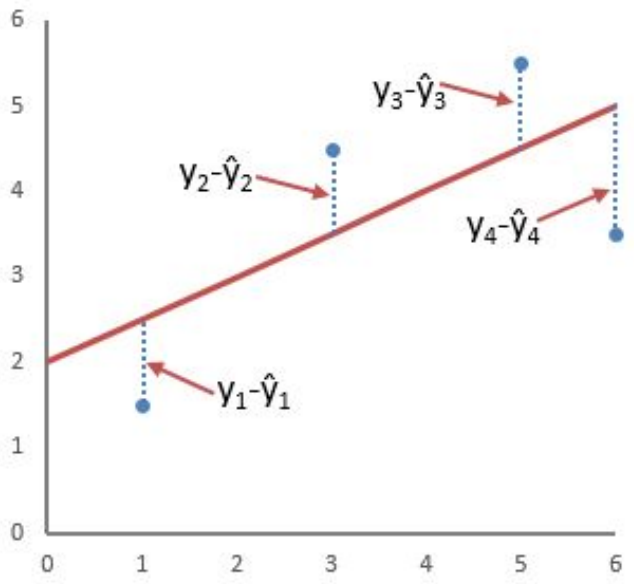
$$\sigma(\bar{b}) = \frac{1}{2}\sum_{k=1}^{M}(Y_k - \hat{Y}_k)^2$$

МНК позволяет получить оценки максимального правдоподобия неизвестных коэффициентов уравнения регрессии при нормальном распределении факторов, но его можно применять и при любом другом распределении.

Регрессионный анализ – метод наименьших квадратов

В основе МНК лежат следующие положения:

1. значения величин ошибок и факторов независимы, а значит, и некоррели-рованы, т.е. предполагается, что механизмы порождения помехи не связаны с механизмом формирования значений факторов;
2. математическое ожидание ошибки должно быть равно нулю, иначе говоря, ошибка является центрированной вели-чиной;
3. выборочная оценка дисперсии ошибки должна быть минимальна.



Изображение отклонения уравнения регрессии от исходных данных

Регрессионный анализ

Качество полученного уравнения регрессии оценивают по степени близости между результатами наблюдений за показателем и предсказанными по уравнению регрессии значениями в заданных точках пространства параметров.

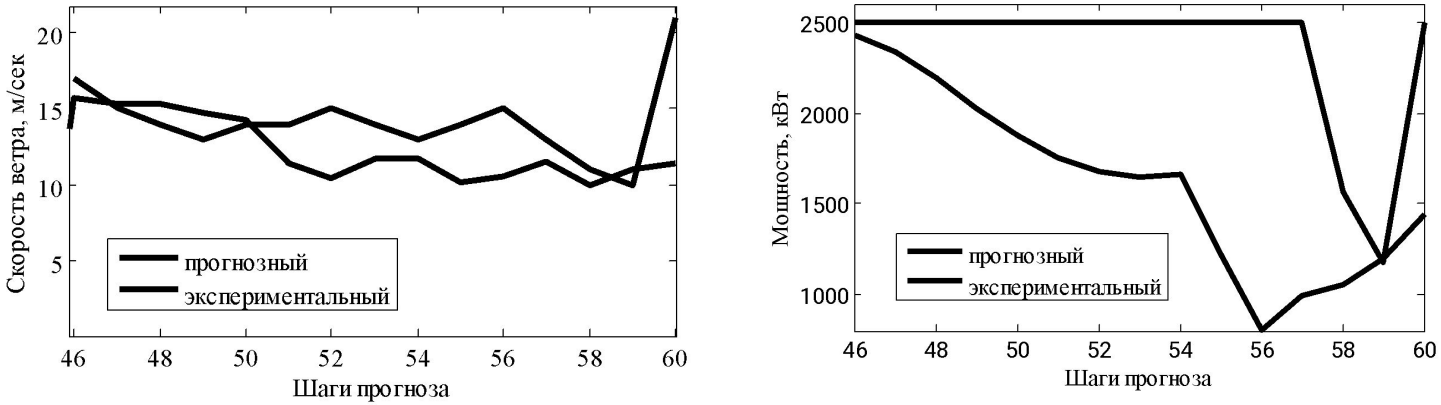
Если результаты близки, то задачу регрессионного анализа можно считать решенной. В противном случае следует изменить уравнение регрессии (выбрать другую степень полинома или вообще другой тип уравнения) и повторить расчеты по оценке параметров. Обычно **применение в уравнениях регрессии полиномов степени выше второй нецелесообразно** .

Главной причиной неточности прогноза является не столько неопределенность экстраполяции линии регрессии, сколько **значительная вариация показателя за счет неучтенных в модели факторов**. Ограничением возможности прогнозирования служит условие стабильности неучтенных в модели параметров и характера влияния учтенных факторов модели. Если резко меняется внешняя среда, то составленное уравнение регрессии потеряет свой смысл.

Пример использования – Прогнозирование выработки мощности ветроустановок на базе регриссионых моделей

Традиционный подход к предсказанию выработки мощности ВЭУ с использованием регрессионных моделей АРПСС заключается в прогнозировании следующего значения ряда, используя известные предыдущие значения ряда. Задача прогнозирования состоит в определении коэффициентов полиномов авторегрессии по данным выборки стационарного процесса выработки мощности ветротурбинами ветростанции, P(t):

$$P(t) = \sum_{j=1}^{P} a_j P_{t-j} + \sum_{j=0}^{q} b_j e_{t-j}$$



Краткосрочное прогнозирование выработки мощности и скорости ветра для ветроустановок для Апшеренского полуострова, Азербайджан

Пример использования – Моделирование распределение температуры оборудования от параметров работы электровозов при их движении

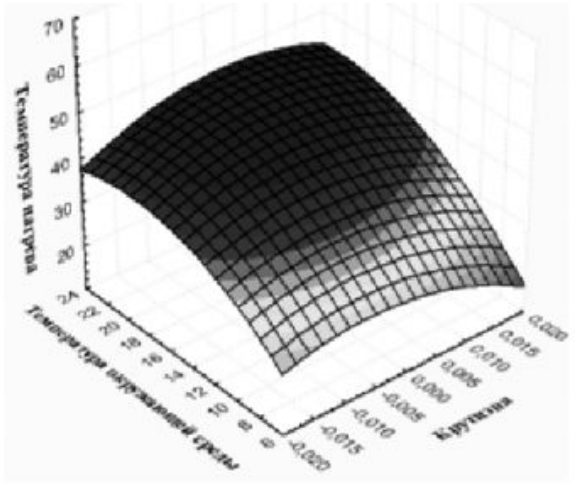
Тепловой нагрев электрооборудования, как правило, определяется большим числом одновременно и совокупно действующих факторов. В связи с этим возникает задача исследования зависимости температуры нагрева (зависимой переменной Y) от нескольких объясняющих переменных X1, X2,… (силы тока, скорости движения, температуры окружающей среды и т.д.). Эта задача решается с помощью множественного регрессионного анализа

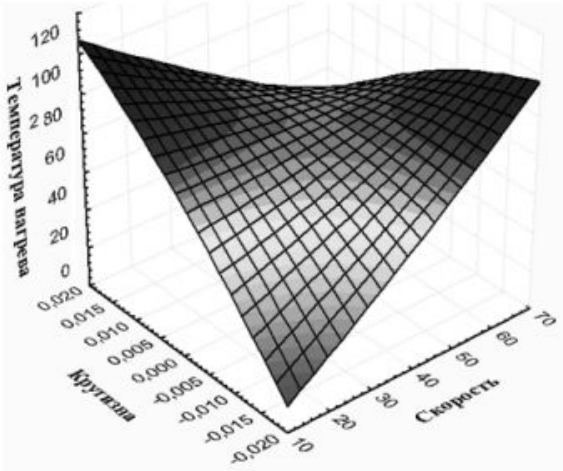
Предложенная методика применена для оценки многофакторного влияния на температуру нагрева различных параметров. В качестве базовых деталей были выбраны силовые шины, поскольку они обладают средним значением теплофизических параметров контролируемых деталей и дают наибольшее число отказов, связанных с перегревом.

Пример использования – Моделирование распределение температуры оборудования от параметров работы электровозов при их движении

Установлено, что максимальный нагрев деталей происходит при увеличении в положительную сторону крутизны профиля пути и маленькой скорости движения поезда, что связано с увеличением силы тока (см. рис.).

Выявлена зависимость возрастания температуры деталей при больших отрицательных значениях крутизны (рекуперация) и большой скорости, так как при увеличении скорости вращения двигателя в режиме генератора возрастает и сила тока, отдаваемая обратно в сеть.





Корреляционный анализ

Для управления сложными системами, на которые воздействует множество факторов, необходимо иметь представление о факторах, влияющих на достижение желаемой от системы или процесса цели. Факторы, влияние которых на объект значительно, должны быть учтены при составлении модели для ее анализа и синтеза управляющей системы. Для принятия решения о включении или исключении какого-либо фактора широко применяется **корреляционный анализ**.

Корреляция *– это статистическая зависимость между случайными величинами, не имеющая строго функционального характера, при которой изменение одной из случайных величин приводит к изменению математического ожидания другой.*

Корреляционный анализ *— метод обработки статистических данных, с помощью которого измеряется теснота связи между двумя или более переменными.*

Корреляционный анализ тесно связан с регрессионным анализом, с его помощью определяют необходимость включения тех или иных факторов в уравнение множественной регрессии, а также оценивают полученное уравнение регрессии на соответствие выявленным связям.

Корреляционный анализ

Употребляется в науке с конца XYIII века. Его ввел французский палеонтолог Жорж Кювье, основавший **"закон корреляции"**, согласно которому череп с рогами обязательно принадлежал травоядному животному, обладавшему копытными конечностями; если же лапа имела когти, то животное было хищным, без рогов, но с крупными клыками.

Жорж Леопоольд Кювье, барон, французский естествоиспытатель, натуралист.

Корреляционный анализ

Корреляция – *это статистическая зависимость между случайными величинами, не имеющая строго функционального характера, при которой изменение одной из случайных величин приводит к изменению математического ожидания другой.*

В статистике принято различать следующие виды корреляции:

- **парная корреляция** – связь между двумя признаками (результативным и факторным, или двумя факторными);
- **частная корреляция** – зависимость между результативным и одним факторным признаками при фиксированном значении других факторных признаков;
- **множественная корреляция** – зависимость результативного и двух или более факторных признаков, включенных в исследование.

Задачей корреляционного анализа является количественное определение тесноты связи между двумя признаками (при парной связи) и между результативным и множеством факторных признаков (при многофакторной связи).

Корреляционный анализ – основная идея

- Идея сопоставления колебаний значений признака относительно друг друга
- Если численные значения одного признака изменяются одновременно со значением другого, то можно предположить, что между ними существует связь
- Следовательно, метод позволяет приблизиться к пониманию причинно-следственных связей

Корреляционный связь

- Характеризует сложный механизм взаимодействия двух или нескольких признаков
- При котором при изменении одного признака случайные варианты второго признака закономерно изменяются
- И величина значений второго признака зависит от величины первого (например, связь между температурой и электрической нагрузкой; социальным статусом и воровством электроэнергии, напряжением и реактивной мощностью т.п.)

Коэффициент корреляции Пирсона

- Предполагает, что:
  - обе переменные распределены нормально
  - связь линейна
- Коэффициент корреляции Пирсона основан на расчете ковариации между двумя перемен-ными:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Корреляционный анализ

- при **r > 0,85** (при этом варьирование признаков взаимосвязано приблизительно на 75% и более) – **весьма тесная связь**,
- при **0,85 > r > 0,7** (при этом взаимосвязанная вариация признаков лежит в пределах 75–50%) – **тесная связь**,
- если **r≤0,7** (при этом варьирование одного признака менее чем на 50% связано с варьированием другого признака) – **связь можно считать слабой.**

Коэффициент Спирмена

- Не предполагает, что данные распределены каким-то особым образом
- Вместо исходных значений использует их ранги
- (!) Интерпретация не настолько проста, как в случае с коэффициентом Пирсона (т.к. связь необязательно линейна)

$$\rho = 1 - \frac{6}{n(n-1)(n+1)} \sum_{i=1}^{n} (R_i - S_i)^2$$

Оценка значимости корреляции

- Оценка коэффициента корреляции, вычисленная по ограниченной выборке, практически всегда отличается от нуля. Но из этого еще не следует, что коэффициент корреляции генеральной совокупности также отличен от нуля.
- Требуется оценить значимость выборочной величины коэффициента или, в соответствии с постановкой задач проверки статистических гипотез, проверить гипотезу о равенстве нулю коэффициента корреляции.
- Если гипотеза H0 о равенстве нулю коэффициента корреляции будет отвергнута, то выборочный коэффициент значим, а соответствующие величины связаны линейным соотношением.

Оценка значимости корреляции

Для проверки гипотезы о значимости коэффициента корреляции используется критерий Стьюдента в виде:

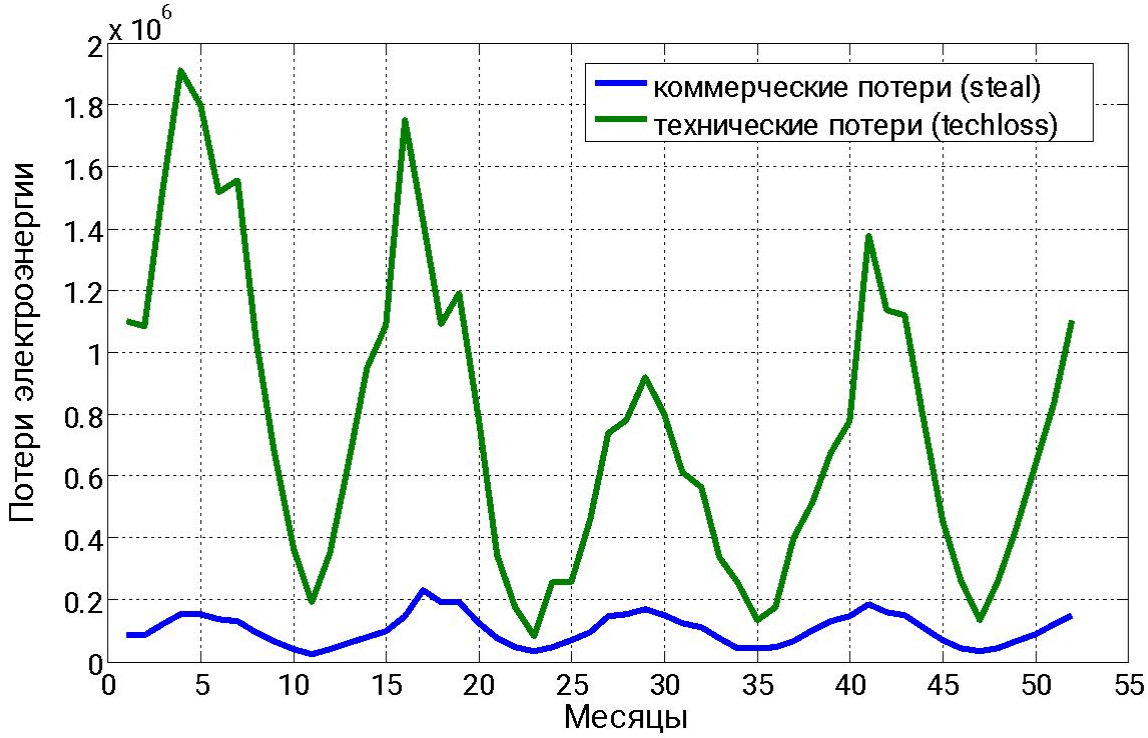
$$t_{набл} = \frac{r_B \sqrt{N-2}}{\sqrt{1-r_B^2}}$$

В этом случае, распределение Стьюдента имеет степень свободы равную.

Проверяемый коэффициент корреляции считается значимым, если значение tнабл по модулю будет больше, чем величина tкр, определенная по таблицам t-распределения

Расчёт коэффициента Пирсона в R

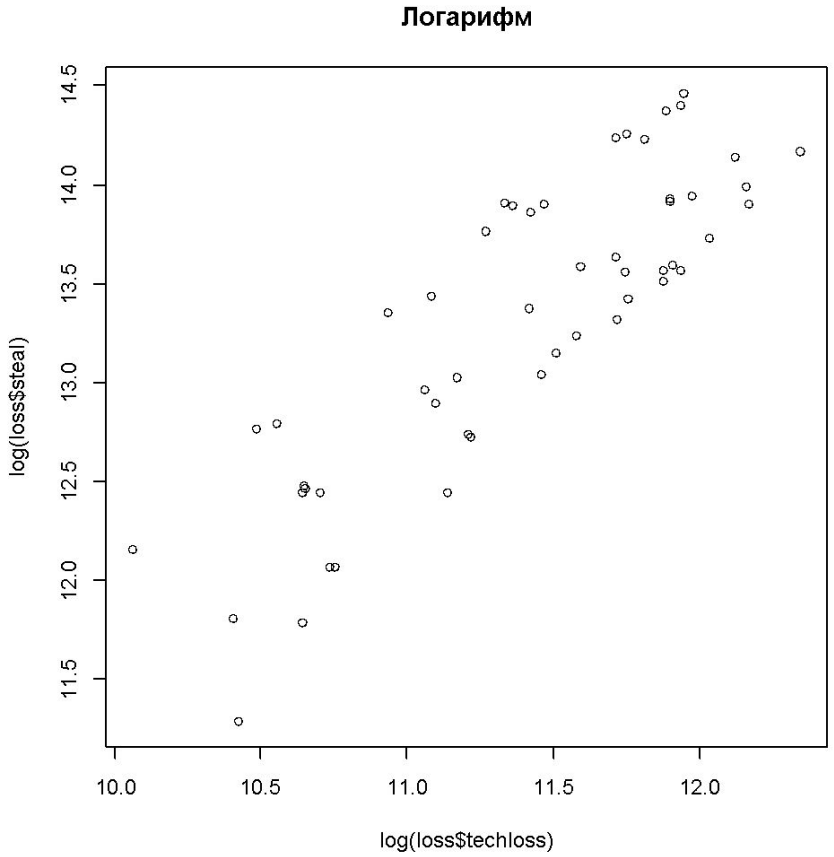
Пример. Даны выборки данных по техническим и коммерческим потерям электроэнергии в электрических сетях г. Братска за 2 года. Необходимо найти коэффициент корреляции между этими параметрами и проверить его статическую значимость

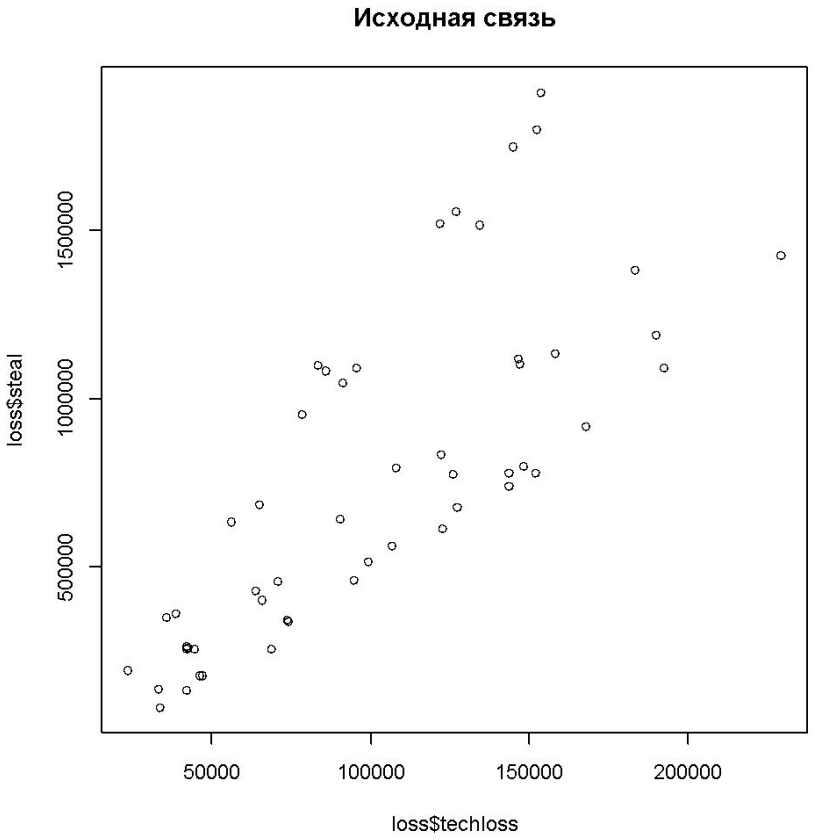


Расчёт коэффициента Пирсона в R

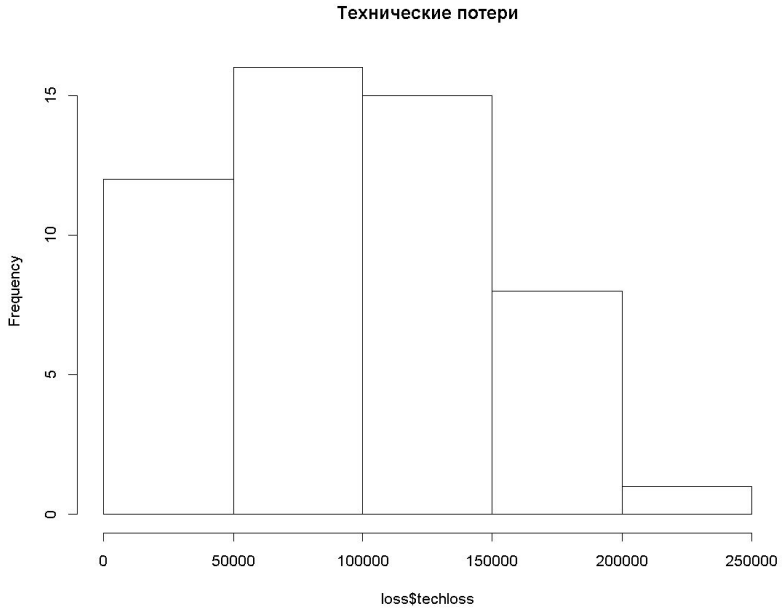
```
< loss <- read.csv ("loss.csv", sep = ";", header=TRUE)

#корреляционный анализ
< cor.test (loss$techloss, loss$steal)

        Pearson's product-moment correlation

data:  loss$techloss and loss$steal
t = 8.4983, df = 50, p-value = 2.848e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6274242 0.8609867
sample estimates:
      cor
0.7687038
```

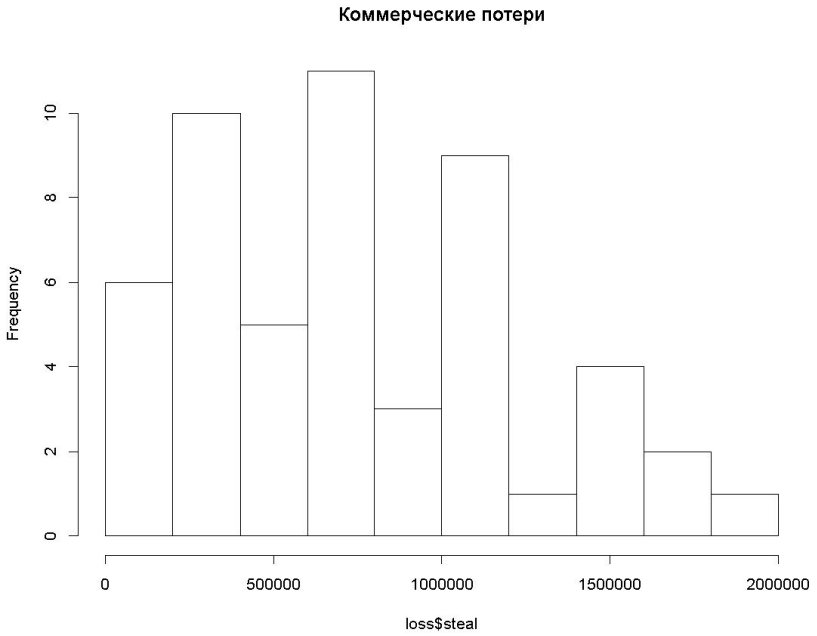
Связь между потерями нелинейна (на исходной шкале)





Ни одна из переменных не распределена нормально






```
Shapiro-Wilk normality test

data: loss$techloss
W = 0.95535, p-value = 0.04928
```

```
Shapiro-Wilk normality test

data:  loss$steal
W = 0.94266, p-value = 0.01438
```

Расчёт коэффициента Спирмена в R

```
#корреляционный анализ по Спирмену
< cor.test (loss$techloss, loss$steal, method = "spearman")

        Spearman's rank correlation rho

data:  loss$techloss and loss$steal
S = 3968, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.8306156
```